

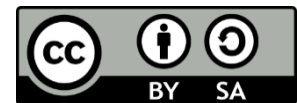
DFG - Projekt

Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow

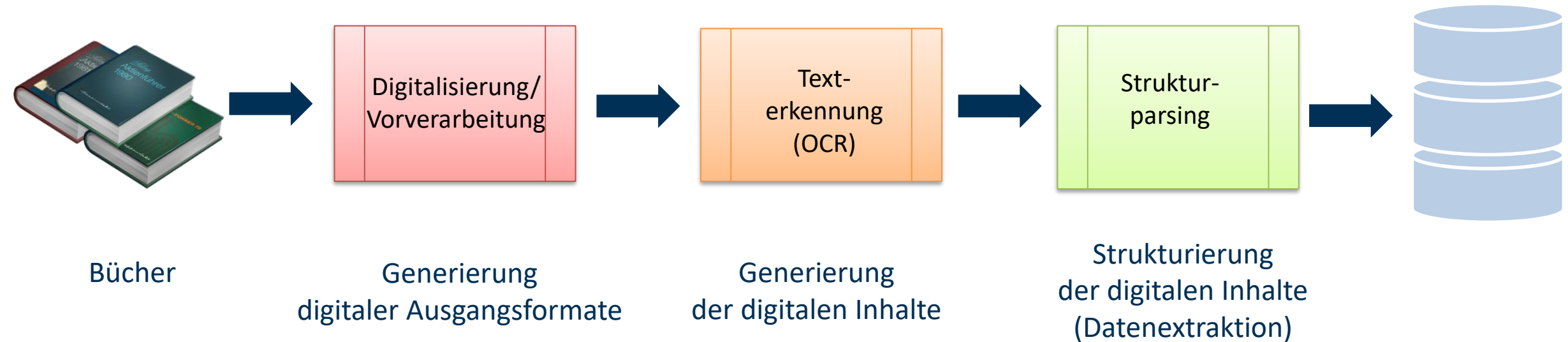


Noah Metzger, Stefan Weil
Universitätsbibliothek Mannheim

30.07.2019



Prozesskette Forschungsdaten aus Digitalisaten



OCR Software (Übersicht)

kommerzielle
Software

fett = eingesetzt in Bibliotheken

ABBYY Finereader

BIT-Alpha

Readiris

OmniPage

Adobe Acrobat

CorelDraw

Microsoft OneNote

...

Tesseract

OCRopus / Kraken /

Calamari

CuneiForm

...

freie Software

ABBYY Cloud OCR

Google Cloud Vision

Microsoft Azure Computer Vision


OCR.space Online OCR ...

Cloud OCR

Tesseract OCR

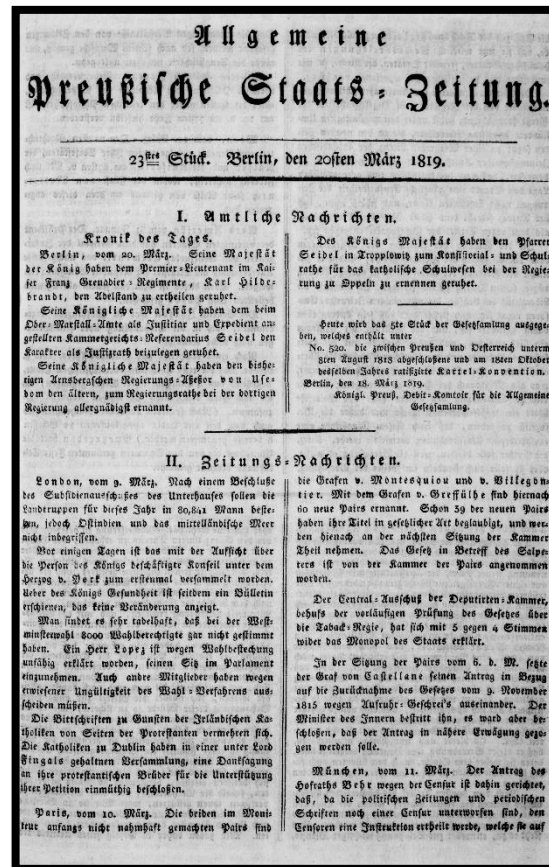
- Open Source
- Komplettlösung „All-in-1“
- Mehr als 100 Sprachen / mehr als 30 Schriften
- Liest Bilder in allen gängigen Formaten (nicht PDF!)
- Erzeugt Text, PDF, hOCR, ALTO, TSV
- Große, weltweite Anwender-Community
- Technologisch aktuell (Texterkennung mit neuronalem Netz)
- Aktive Weiterentwicklung u. a. im DFG-Projekt OCR-D

Tesseract an der UB Mannheim

- Verwendung im DFG-Projekt „Aktienführer“
<https://digi.bib.uni-mannheim.de/aktienfuehrer/>
- Volltexte für Deutscher Reichsanzeiger und Vorgänger
<https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger>
- DFG-Projekt „OCR-D“ <http://www.ocr-d.de/>,  OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung
von Verfahren der Optical Character Recognition (OCR)
Modulprojekt „Optimierter Einsatz von OCR-Verfahren – Tesseract als
Komponente im OCR-D-Workflow“:
Schnittstellen, Stabilität, Performance und praktische Einsetzbarkeit,
Erweiterungen wie z. B. Konfidenzen

Übersicht

Digitalisate für Forschungsdaten an der UB Mannheim



Reichsanzeiger

Berlin, Hauptstadt der DDR		
a) Übersicht der Stadtbezirke		
Stadtbezirksnummer	Stadtbezirke	Zahl der Ortsteile
1501	Berlin-Mitte	—
1504	Berlin-Prenzlauer Berg	—
1505	Berlin-Friedrichshain	—
1509	Berlin-Marzahn	5
1515	Berlin-Treptow	7
1516	Berlin-Köpenick	11
1517	Berlin-Lichtenberg	3
1518	Berlin-Weißensee	5
1519	Berlin-Pankow	11

b) Ortsteile nach Stadtbezirken		
Stadtbezirke Ortsteile	Stadtbezirksnummer	Ortsteilnummer
Berlin-Mitte	1501	
Berlin-Prenzlauer Berg	1504	
Berlin-Friedrichshain	1505	
Berlin-Marzahn	1509	
Berlin-Marzahn	150900	01
Berlin-Biesdorf	150900	02
Berlin-Kaustdorf	150900	03
Berlin-Mahlsdorf	150900	04
Berlin-Hellersdorf	150900	05
Berlin-Treptow	1515	
Berlin-Adlershof	151500	01
Berlin-Altglienicke	151500	02
Berlin-Baumshulenberg	151500	03
Berlin-Bohnsdorf	151500	04
Berlin-Johannisthal	151500	05
Berlin-Niederschöneweide	151500	06
Berlin-Treptow	151500	08
Berlin-Köpenick	1516	
Berlin-Friedrichshagen	151600	02
Berlin-Grünau	151600	03
Berlin-Karolinenhof	151600	04
Berlin-Köpenick	151600	05
Berlin-Müggelheim	151600	06
Berlin-Oberschöneweide	151600	07
Berlin-Rahnsdorf	151600	08
Berlin-Hessenwinkel	151600	10

Gemeindeverzeichnisse

A

EDUARD AHLBORN AKTIENGESellschaft

Sitz: 32 Hildesheim, Lüntzelstraße 22,
Postfach 530

Fernruf: Sa.-Nr. 8 32 71-75

Fernschreiber: 09 2763

Vorstand:

Ernst Morach, Hildesheim, Vors.;
Dr. phil. Karl Bechold, Hildesheim

Aufsichtsrat:

Ernst Morach, Hannover, Vors.;
Dr. Werner Anders, Hannover, stellv.
Vors.;
Justus Mundi, Freudenberg-Siegen;
Professor Dr.-Ing. Eduard Pestel, Han-
nover;

Achim Seibert, Bernried;
Bernd Wagner, Hildesheim;
Arbeitnehmervertreter:

Franz Atonhan, Hildesheim;
Theodor Mannes, Borsum;
Walter Mundry, Hildesheim

Gründung: 1927

Tätigkeitsgebiet:

Fabrikation und Vertrieb von Molkerei-
und Kältemaschinen, Blechwaren, Ma-
schinen und Geräten für das Nahrungs-
mittelgewerbe sowie der Handel mit die-
sen Gegenständen und in Bedarfartikeln
aller Art für das Nahrungsmittelgewerbe,
ferner der Vertrieb von landwirtschaftli-
chen Maschinen und Geräten; Betrieb so-
stiger industrieller und Handelsunterneh-
mungen.

Geschäftsjahr: Kalenderjahr

Stimmrecht d. Aktien i. d. H.-V.:
Je nom. DM 1 000,- = 1 Stimme

Zahlstellen:

Gesellschaftskasse, Hildesheim;
Deutsche Bank AG, Hannover und Hildes-
heim;
Hallbaum, Maler & Co., Hannover

Grundkapital: DM 3 000 000,-
Umstellung 1:1 durch H.-V.v. 1.11.1950.

Börsennotiz: Hannover (Freiv.)

Wertpapier-Kenn-Nr.: 500980

Stückelung:
3 000 Inh.-St.-Akt. zu je DM 1 000,-

Großaktionär: Familienbesitz
(ca. 60 %);

Rest Streubesitz

Aktienkurse (Hannover):

Notierung seit 9.2.1955

ultimo 1955 130 % +)

" 1956 128 %

" 1957 136 %

" 1958 185 %

" 1959 355 %

" 1960 570 %

" 1961 370 %

" 1962 301 %

30. Sept. 1963 341 %

+) ab Tag der Notierung Kurs für DM-
Nennwert

Dividenden auf Stammaktien:

1/1948/49-1958: insgesamt 59 %

1959: 12 % (Div. Sch. Nr. 6)

1960: 13 % (Div. Sch. Nr. 7)

1961 u. 1962: je 12 % + 2 % Bonus

(Div. Sch. Nr. 8 u. 9)

Aus den Bilanzen

	31.12.1961	31.12.1962
	(in 1 000 DM)	
Anlagevermögen	4 384	4 229
Umlaufvermögen	13 699	13 645
(darunter)		
Vorräte	7 949	8 000
Lieferantenverbindungen	4 634	4 671
Barmittel einschl. Wertpapiere	253	359
Eigenkapital	4 171	4 018
(davon A.-K.)	3 000	3 000
Fremdkapital	13 360	13 313
(darunter)		
Anzahlungen	1 760	1 387
Gewinn nach Vortrag	431	437

Aus den Gewinn- und Verlust-
rechnungen

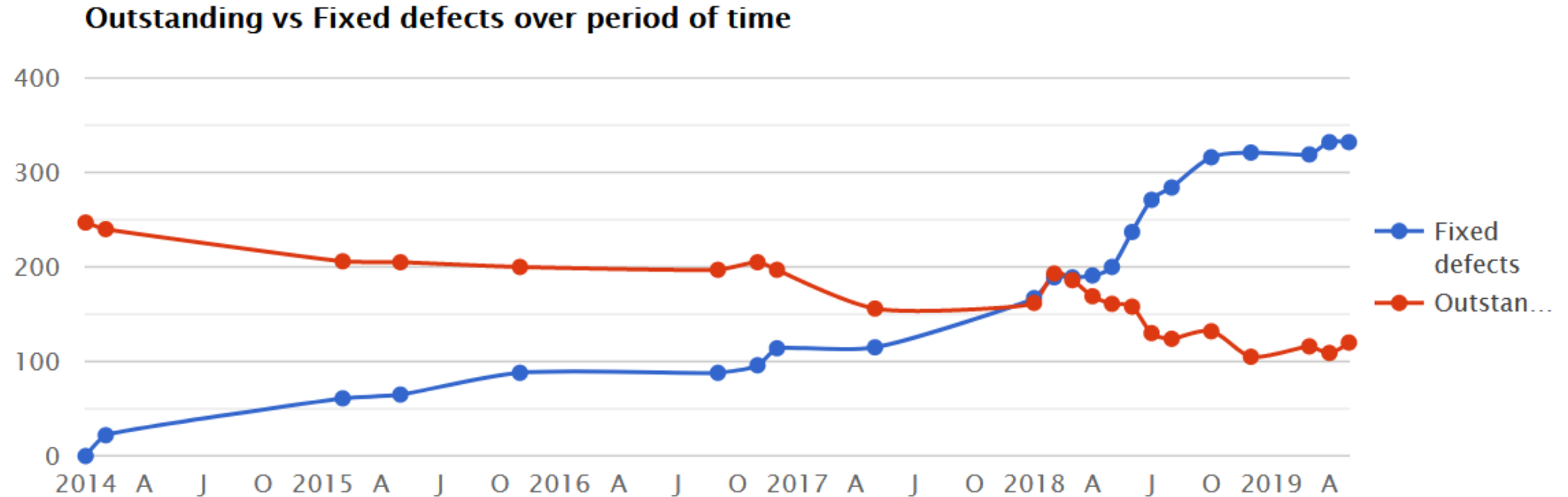
	1961	1962
Grundkapital: DM 3 000 000,-		
Löhne und Gehälter	8 504	9 475
Abschreibungen	722	632
Besitzsteuern	776	410
Sonstige Steuern	998	1 060
Umsatzerlöse	37 162	36 597

17

Aktienführer

DFG-Projekt „OCR-D“

Stabilitätsverbesserung



DFG-Projekt „OCR-D“

Performanceverbesserung

- Etwa **90 %** der verwendeten Rechenzeit wird für Skalarprodukte aufgewendet
- Verwendung von 32-bit Werten anstelle der ursprünglich verwendeten 64-bit Werten
- Nutzung des Kahan Summations Algorithmus um den entstehenden Verlust an Genauigkeit zu kompensieren

DFG-Projekt „OCR-D“

Performanceverbesserung

- Durchschnittliche Zeitersparnis von **42,5 %**
- Durchschnittliche Performanceverbesserung von **74 %**
- Trotz geringerer Genauigkeit **keine** Abweichung der Ergebnisse

DFG-Projekt „OCR-D“

Schnittstellen für andere Modul Projekte

- Geplante OCR post-Korrektur der Universität Leipzig
- Bereitstellung eines Ausgabemodus, welcher zusätzliche Zeichen - Alternativen zu dem bestehenden Ergebnis liefert.



DFG-Projekt „OCR-D“

Schnittstellen für andere Modul Projekte

本



Text-
erkennung
(OCR)

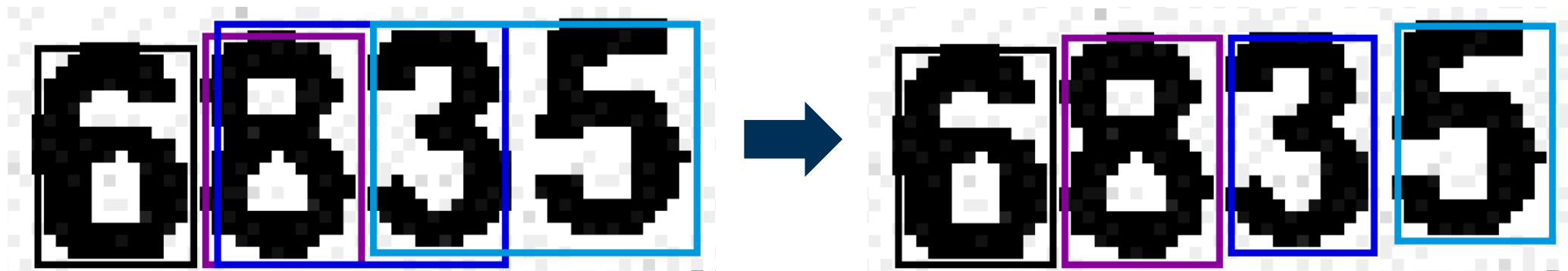


```
box 22 16 254 235; x_wconf 94'>本  
1_1'>  
title='x_confs 98.085274'>本</span>  
title='x_confs 41.138905'>人</span>  
title='x_confs 32.802567'>末</span>  
title='x_confs 15.717819'>な</span>  
title='x_confs 1.7657089'>ネ</span>
```

DFG-Projekt „OCR-D“

Positive Nebeneffekte

- Allgemeine Anwendbarkeit der Schnittstelle
- Wiederverwendung von Teilalgorithmen zur Erstellung besserer Begrenzungsrahmen im Standardprozess von Tesseract



Ein Blick **zurück** und nach **vorne**

Derzeitiger Stand:

- Alle Tools als **Open Source** öffentlich auf GitHub:
<https://github.com/tesseract-ocr/tesseract>



Ausblick:

- Kooperationsprojekt **OCR-BW** (gemeinsam mit Tübingen, Landesprojekt BW, gestartet)
 - Aufbau eines Kompetenzzentrums Volltexterkennung für Bibliotheken und Archive
- DFG-Projekt **Reichsanzeiger** (bewilligt)
- DFG-Projekt **OCR-D** Folgeprojekt (offen)

Literatur

- Weil, S., & Zumstein, P. (2016). Mit freier Software Text in Digitalisaten erkennen. <https://speakerdeck.com/zuphilip/mit-freier-software-text-in-digitalisaten-erkennen-ocr-praxis-an-der-ub-mannheim>
- Baierer, K., & Zumstein, P. (2016). Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture*, 4(2), 72-83. <https://doi.org/10.12685/027.7-4-2-155>
- Kamlah, J., Stegmüller, J. (2018). Ocromore – Combining multiple OCR-engine results to improve character recognition accuracy. <https://zenodo.org/record/1493860>
- Kamlah, J., Stegmüller, J., Schumm, I., Zumstein, P. (2019). Automatisierte Optimierung und Strukturierung von OCR-Ergebnissen mit nachnutzbaren Werkzeugen. <https://ub-madoc.bib.uni-mannheim.de/48940>
- Weil, S. (2019). Vom Bild zum Text. Automatisierte Texterkennung in historischen Drucken mit der freien Software Tesseract. <https://nbn-resolving.org/urn:nbn:de:0290-opus4-163511>

Bildquellen

- Titelseite:
<https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126884/>
<https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126880/>
<https://pixabay.com/de/vectors/werkzeug-schraubenschl%C3%BCssel-3456474/>
<https://commons.wikimedia.org/wiki/File:Opensource.svg>
- DFG-Logo: <https://www.dfg.de/>
- GitHub Logos: <https://github.com/logos>
- OCR-D Logo: <http://www.ocr-d.de/>
- Universität Leipzig Logo: https://de.wikipedia.org/wiki/Universit%C3%A4t_Leipzig